BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

Translated Excerpt of Citation 3

Japanese Patent Laid-Open Publication No. HEI 5-219932

* * * * * * * * * * * * * * * * *

Paragraphs [0015] – [0021]:

"[0015]

[Action]

According to the genetic information examining device 1 of the present invention shown in Fig. 1(a), when character sequence expression of the amino acid arrangement for examination is "ABCBDAB" while character sequence expression of the amino acid arrangement for comparison is "BDCABA", for example, detecting means 10detects that the number of the longest common characters in these two character sequences is "4", and calculating means 11, in response to the detection result, calculates the rate value of 57% (=4/7), when based on the character sequence length of the amino acid arrangement for examination, and when based on the character sequence length of the amino acid arrangement for comparison, the rate value of 67% (=4/6) is calculated.

[0016]

The output controlling means 12 outputs the calculated rate value to an outputting device 2 to notify a user of the similarity between the amino acid arrangement for examination and the amino acid arrangement for comparison. On the other hand, evaluating means 13 compares the calculated rate value with the regular standard to evaluate the similarity between the amino acid arrangement for examination and the amino acid arrangement for comparison mechanically, and to perform the above-mentioned homology examination.

[0017]

As described above, the genetic information examining device 1 of the present invention according to Fig. 1 (a) has the composition which calculates the number of the longest common characters of the amino acid arrangement for examination expressed in a character sequence and the amino acid arrangement for comparison expressed in a character sequence. Since it is the composition of evaluating the similarity between the amino acid arrangement for examination and the amino acid arrangement for comparison according to this number of the longest common characters, compared with

a comparison method of the amino acid arrangement by dynamic programming technique, the similarity of two amino acid arrangements can be evaluated with small memory capacity and at high speed.

[0018]

According to the genetic information examining device 1 of the present invention shown in Fig. 1(b), when character sequence expression of the amino acid arrangement for examination is "ABCBDAB" while character sequence expression of the amino acid arrangement for comparison is "BDCABA", detecting means 20detects that the longest common portion sequences in these two character sequences are "BDAB", "BCBA", "BDAB" and "BCAB". In response to the detection result, specifying means 21 specifies the arrangement position of each longest common portion sequence in the amino acid arrangement for examination and the amino acid arrangement for comparison, and also specifies the character sequence length existing between the arrangement positions.

[0019]

Then, the output controlling means 22 outputs the longest common portion sequence detected by the detecting means 20 as it is, outputs the character sequence length specified by the specifying means 21 so that it is corresponding to the longest common portion sequence, and moreover, outputs these two amino acid arrangements while aligning so that the longest common portion sequences of the amino acid arrangement for examination and the amino acid arrangement for comparison are corresponding to each other, according to the arrangement position specified by the specifying means 21. In such a manner, the output controlling means 22 notifies the user of the similarity between the amino acid arrangement for examination and the amino acid arrangement for comparison.

[0020]

On the other hand, when the evaluating means 23 evaluates whether the amino acid arrangement for comparison is included in the amino acid arrangement for examination in case the amino acid arrangement for comparison is expressed in the successive character sequence or in character sequence containing non-specified character sequence between arrangement positions, the evaluating means 23 evaluates mechanically whether the longest common portion sequence detected by detecting means 20 and the amino acid arrangement for comparison coincide or not, taking into consideration the

specified result by the specifying means 21, and performs the above-described motif examination.

[0021]

As described above, the genetic information examining device 1 of the present invention according to Fig. 1 (b) has the composition specifying the longest common portion sequence of the amino acid arrangement for examination expressed in a character sequence and the amino acid arrangement for comparison expressed in a character sequence. Since it is the composition of evaluating the similarity between the amino acid arrangement for examination and the amino acid arrangement for comparison according to the longest common portion sequence, compared with a comparison method of the amino acid arrangement by dynamic programming technique, the similarity of two amino acid arrangement can be evaluated with small memory capacity and at high speed. "

* * * * * * * * * * * * * * * * * * *

Citation 3

(19)日本国特許庁（ＪＰ）　(12)**公開特許公報**（Ａ）　(11)特許出願公開番号

特開平5－219932

(43)公開日　平成５年(1993)８月31日

(51)Int.Cl.[5]		識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｃ１２Ｍ	1/00		Ａ　9050－4Ｂ		
Ｇ０６Ｆ	15/40	５１０	Ｅ　7060－5Ｌ		
	15/42		Ａ　7060－5Ｌ		

審査請求　未請求　請求項の数５(全 17 頁)

(21)出願番号　特願平4－21012

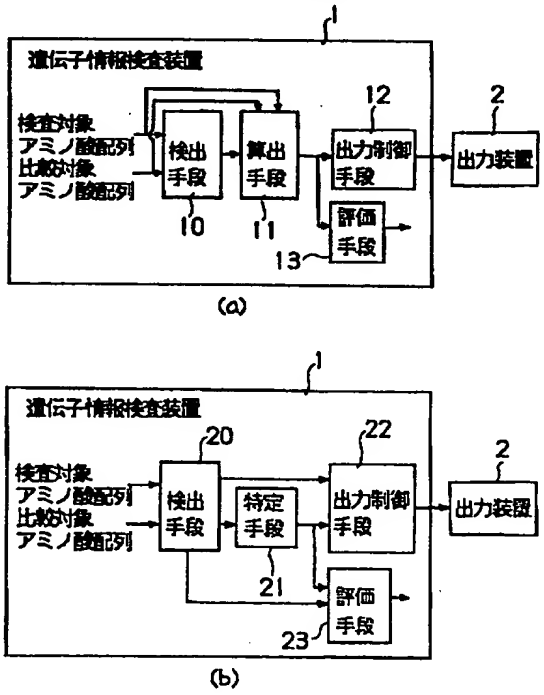
(22)出願日　平成４年(1992)２月６日

(71)出願人　000005223
富士通株式会社
神奈川県川崎市中原区上小田中1015番地
(72)発明者　大矢　真弓
神奈川県川崎市中原区上小田中1015番地
富士通株式会社内
(72)発明者　相川　聖一
神奈川県川崎市中原区上小田中1015番地
富士通株式会社内
(74)代理人　弁理士　森田　寛　(外１名)

(54)【発明の名称】　遺伝子情報検査装置

(57)【要約】
【目的】本発明は、検査対象のアミノ酸配列と、比較対象のアミノ酸配列との間の類似性を評価する遺伝子情報検査装置に関し、簡単な処理機構に従って類似性を評価することのできるようにすることを目的とする。
【構成】アミノ酸を文字で表現する構成を採り、かつ、文字列で表現される検査対象アミノ酸配列と、文字列で表現される比較対象アミノ酸配列との最長共有文字数を検出する検出手段１０と、この検出手段１０により検出される最長共有文字数と、検査対象アミノ酸配列又は比較対象アミノ酸配列の文字列長との割合を算出する算出手段１１とを備えるように構成し、また、アミノ酸を文字で表現する構成を採り、かつ、文字列で表現される検査対象アミノ酸配列と、文字列で表現される比較対象アミノ酸配列との最長共有部分列を検出する検出手段(20)を備えるように構成する。

本発明の原理構成図



1

【特許請求の範囲】

【請求項１】　検査対象のアミノ酸配列と、比較対象のアミノ酸配列との間の類似性を評価する遺伝子情報検査装置において、

アミノ酸を文字で表現する構成を採り、

かつ、文字列で表現される検査対象アミノ酸配列と、文字列で表現される比較対象アミノ酸配列との最長共有文字数を検出する検出手段(10)と、

上記検出手段(10)により検出される最長共有文字数と、検査対象アミノ酸配列又は比較対象アミノ酸配列の文字列長との割合を算出する算出手段(11)とを備えることを、

特徴とする遺伝子情報検査装置。

【請求項２】　検査対象のアミノ酸配列と、比較対象のアミノ酸配列との間の類似性を評価する遺伝子情報検査装置において、

アミノ酸を文字で表現する構成を採り、

かつ、文字列で表現される検査対象アミノ酸配列と、文字列で表現される比較対象アミノ酸配列との最長共有部分列を検出する検出手段(20)を備えることを、

特徴とする遺伝子情報検査装置。

【請求項３】　請求項２記載の遺伝子情報検査装置において、

検査対象アミノ酸配列及び比較対象アミノ酸配列の持つ最長共有部分列の配列位置を特定する特定手段(21)を備えることを、

特徴とする遺伝子情報検査装置。

【請求項４】　請求項３記載の遺伝子情報検査装置において、

出力装置に対して２つのアミノ酸配列を出力する構成を採って、この出力時に、特定手段(21)により特定される配列位置に従って、この２つのアミノ酸配列の持つ最長共有部分列が対応付けられるべく該アミノ酸配列をアラインメントして出力していくよう処理することを、

特徴とする遺伝子情報検査装置。

【請求項５】　請求項３記載の遺伝子情報検査装置において、

出力装置に対して２つのアミノ酸配列の持つ最長共有部分列を出力する構成を採って、この出力時に、特定手段(21)により特定される配列位置に従って規定される該配列位置間の持つ文字列長を、該最長共有文字列に対応付けて出力していくよう処理することを、

特徴とする遺伝子情報検査装置。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は、検査対象のアミノ酸配列と、比較対象のアミノ酸配列との間の類似性を評価する遺伝子情報検査装置に関し、特に、簡単な処理機構に従って類似性を評価することのできる遺伝子情報検査装置に関するものである。

2

【０００２】医薬品の開発等に必要なタンパク質工学の分野では、分子生物学の進歩に伴って大量に遺伝情報が蓄積され始め現在データベース化が進んでいる。これらの大量に蓄積された遺伝情報から、タンパク質の構造や機能等の生物学的に意味のある情報を抽出していくことが要求されている。この抽出処理は、高速処理を可能にするためにも、できる限り簡単な処理機構により実現していくことが好ましいのである。

【０００３】

【従来の技術】遺伝子の本体はＤＮＡであり、Ａ（アデニン）、Ｔ（チミン）、Ｃ（シトシン）、Ｇ（グアニン）という４つの塩基で構成される塩基配列として表現される。また、生体を構成するアミノ酸は約２０種類あり、これまでに、塩基配列中の３つの塩基の並びと各アミノ酸とが対応することが解明されている。従って、生体内では、ＤＮＡの塩基配列に従ってアミノ酸が合成され、合成されたアミノ酸が折り畳まれることによってタンパク質が形作られることになる。

【０００４】上述したように、分子生物学の発展に伴い、塩基やアミノ酸の配列の決定法が確立したことによって、塩基配列データ、アミノ酸配列データ等の遺伝情報が大量に蓄積され始めている。このため、遺伝子情報処理の分野では、蓄積された膨大な遺伝情報の中から、タンパク質の構造や機能等に関する生物情報をいかにして抽出するかが中心課題となってきた。

【０００５】このような生物情報を抽出する際の基本的手法は、アミノ酸の配列を比較することである。これは、アミノ酸の配列が類似していることは、生物学的機能にも類似があると考えられているためである。

【０００６】このようなことを背景にして、評価対象のアミノ酸配列の機能を推定するために、機能が解明されている既知のアミノ酸配列データベースから、評価対象のアミノ酸配列と類似するアミノ酸配列を検索する相同性探索や、比較するアミノ酸配列間での違いと類似性が明確になるようにアミノ酸配列を並び変えるアラインメントが行われるようになってきている。

【０００７】また、アミノ酸配列の中で生物にとって重要な機能をコードしている領域は、進化の過程でも保存されていると考えられている。例えば、異なる生物種で同じ機能を持つタンパク質のアミノ酸配列を比較すると、共通に存在する配列パターンがあることが知られている。このような配列パターンはモチーフと呼ばれている。これから、アミノ酸配列中にどのようなモチーフが含まれているかを調べることによって、タンパク質の性質や機能を解明することができるだけでなく、既存のタンパク質に対する強化、機能の付加、新しいタンパク質の合成等、多岐に渡ってタンパク質工学の分野に応用することができる。これから、モチーフを検索することが行われるようになってきている。

【０００８】従来、２つのアミノ酸配列を比較する方法

としては、音声認識処理等で用いられているダイナミックプログラミング手法が用いられている。

【０００９】

【発明が解決しようとする課題】しかしながら、ダイナミックプログラミング手法によるアミノ酸配列の比較方法では、２次元的にアミノ酸配列を比較していくために、大きなメモリ容量が必要になるとともに、処理時間も多くかかるという問題点があった。

【００１０】本発明はかかる事情に鑑みてなされたものであって、検査対象のアミノ酸配列と、比較対象のアミノ酸配列との間の類似性を簡単な処理機構に従って評価することのできる新たな遺伝子情報検査装置の提供を目的とするものである。

【００１１】

【課題を解決するための手段】図１（ａ）に本発明の第１の発明の原理構成、図１（ｂ）に本発明の第２の発明の原理構成を図示する。

【００１２】図１（ａ）(ｂ）中、１は本発明を具備する遺伝子情報検査装置であって、アミノ酸を文字で表現する構成を採って、検査対象のアミノ酸配列と、比較対象のアミノ酸配列との間の類似性を評価するもの、２は遺伝子情報検査装置１に接続される出力装置である。

【００１３】図１（ａ）に従う本発明の遺伝子情報検査装置１は、文字列で表現される検査対象アミノ酸配列と、文字列で表現される比較対象アミノ酸配列との最長共有文字数を検出する検出手段１０と、検出手段１０により検出される最長共有文字数と、検査対象アミノ酸配列又は比較対象アミノ酸配列の文字列長との割合を算出する算出手段１１と、算出手段１１の算出する割合値を出力装置２に出力する出力制御手段１２と、算出手段１１の算出する割合値に従って、検査対象アミノ酸配列と比較対象アミノ酸配列との間の類似性を評価する評価手段１３とを備える。

【００１４】図１（ｂ）に従う本発明の遺伝子情報検査装置１は、文字列で表現される検査対象アミノ酸配列と、文字列で表現される比較対象アミノ酸配列との最長共有部分列を検出する検出手段２０と、検査対象アミノ酸配列及び比較対象アミノ酸配列の持つ最長共有部分列の配列位置を特定するとともに、その特定結果に従って、その配列位置間に存在する文字列長を特定する特定手段２１と、検出手段２０の検出結果や特定手段２１の特定結果を出力装置２に出力する出力制御手段２２と、検出手段２０の検出結果や特定手段２１の特定結果に従って、検査対象アミノ酸配列と比較対象アミノ酸配列との間の類似性を評価する評価手段２３とを備える。

【００１５】

【作用】図１（ａ）に従う本発明の遺伝子情報検査装置１では、例えば、検査対象アミノ酸配列の文字列表現が“ＡＢＣＢＤＡＢ”で、比較対象アミノ酸配列の文字列表現が“ＢＤＣＡＢＡ”である場合に、検出手段１０は、この２つの文字列での最長共有文字数が“４”であることを検出し、算出手段１１は、この検出結果を受けて、検査対象アミノ酸配列の文字列長を基準にする場合には５７％（＝４÷７）という割合値を算出し、比較対象アミノ酸配列の文字列長を基準にする場合には６７％（＝４÷６）という割合値を算出する。

【００１６】そして、出力制御手段１２は、この算出された割合値を出力装置２に出力していくことで、ユーザに対して、検査対象アミノ酸配列と比較対象アミノ酸配列との間の類似性の評価値を通知し、一方、評価手段１３は、この算出された割合値を規定の基準値と比較することで、検査対象アミノ酸配列と比較対象アミノ酸配列との間の類似性を機械的に評価して上述の相同性検査を実行していく。

【００１７】このように、図１（ａ）に従う本発明の遺伝子情報検査装置１は、文字列で表現される検査対象アミノ酸配列と、文字列で表現される比較対象アミノ酸配列との最長共有文字数を算出する構成を採って、この最長共有文字数に従って、検査対象アミノ酸配列と比較対象アミノ酸配列との間の類似性を評価する構成を採るものであることから、ダイナミックプログラミングム手法によるアミノ酸配列の比較方法に比べて、小さなメモリ容量で、かつ高速に２つのアミノ酸配列の類似性を評価することができるのである。

【００１８】図１（ｂ）に従う本発明の遺伝子情報検査装置１では、例えば、検査対象アミノ酸配列の文字列表現が“ＡＢＣＢＤＡＢ”で、比較対象アミノ酸配列の文字列表現が“ＢＤＣＡＢＡ”である場合に、検出手段２０は、この２つの文字列での最長共有部分列が“ＢＤＡＢ”、“ＢＣＢＡ”、“ＢＤＡＢ”、“ＢＣＡＢ”であることを検出する。この検出結果を受けて、特定手段２１は、検査対象アミノ酸配列及び比較対象アミノ酸配列の持つ各最長共有部分列の配列位置を特定するとともに、その配列位置間に存在する文字列長を特定する。

【００１９】そして、出力制御手段２２は、検出手段２０の検出する最長共有部分列をそのまま出力したり、特定手段２１により特定される文字列長をこの最長共有部分列に対応付けて出力したり、特定手段２１により特定される配列位置に従って、検査対象アミノ酸配列及び比較対象アミノ酸配列の持つ最長共有部分列が対応付けられるべくこの２つのアミノ酸配列をアラインメントして出力したりしていくことで、ユーザに対して、検査対象アミノ酸配列と比較対象アミノ酸配列との間の類似性を通知する。

【００２０】一方、評価手段２３は、比較対象アミノ酸配列が連続する文字列で表現される場合や、規定されない文字列を配列位置間に含む文字列のもので表現される場合にあって、検査対象アミノ酸配列にこの比較対象アミノ酸配列が含まれているか否かを評価していくときには、特定手段２１の特定結果を考慮しつつ、検出手段２

０の検出する最長共有部分列と比較対象アミノ酸配列とが一致するか否かを機械的に評価していくことで上述のモチーフ検査を実行していく。

【００２１】このように、図１（ｂ）に従う本発明の遺伝子情報検査装置１は、文字列で表現される検査対象アミノ酸配列と、文字列で表現される比較対象アミノ酸配列との最長共有部分列を特定する構成を採って、この最長共有部分列に従って、検査対象アミノ酸配列と比較対象アミノ酸配列との間の類似性を評価する構成を採るものであることから、ダイナミックプログラミング手法によるアミノ酸配列の比較方法に比べて、小さなメモリ容量で、かつ高速に２つのアミノ酸配列の類似性を評価することができるのである。

【００２２】

【実施例】以下、実施例に従って本発明を詳細に説明する。図２に、本発明を実装する遺伝子情報検査装置１の一実施例を図示する。図中、４０は遺伝子情報検査装置１に接続される入力装置、４１は入力装置４０の備えるキーボードやマウス等の対話装置、４２は遺伝子情報検査装置１に接続されるディスプレイ装置、５０は文字列で表現されるアミノ酸配列情報を管理するアミノ酸配列データベース、６０は文字列で表現されるモチーフ配列情報を管理するモチーフデータベースである。

【００２３】この実施例の遺伝子情報検査装置１は、入力装置４０から入力されてくるアミノ酸配列の文字列と、アミノ酸配列データベース５０やモチーフデータベース６０から与えられるアミノ酸配列の文字列との間の最長共有文字数や、最長共有部分列（ＬＣＳ：longest common subsequence）や、最長共有部分列の展開位置を検出するＬＣＳ検出部３０と、ＬＣＳ検出部３０の結果に従って、ＬＣＳ検出部３０の検出対象となった２つのアミノ酸配列の相同性を判定する相同性判定部３１と、相同性判定部３１の検出結果に従って、入力装置４０から入力されてくるアミノ酸配列と相同なアミノ酸配列をアミノ酸配列データベース５０から検索する相同性探索部３２と、ＬＣＳ検出部３０の結果に従って、入力装置４０から入力されてくるアミノ酸配列と相同なモチーフ配列をモチーフデータベース６０から検索するモチーフ探索部３３と、ＬＣＳ検出部３０の検出結果に従って、入力装置４０から入力されてくるアミノ酸配列の文字列と、アミノ酸配列データベース５０やモチーフデータベース６０から与えられるアミノ酸配列の文字列とをアラインメントするアラインメント部３４と、各処理部の処理結果をディスプレイ装置４２に表示する表示部３５とを備える。

【００２４】次に、図３ないし図５に示す処理フローに従って、ＬＣＳ検出部３０の実行する処理について詳細に説明する。ここで、図３に示す処理フローは、検査対象となる２つのアミノ酸配列の持つ最長共有文字数を検出するための処理フローであり、図４及び図５に示す処理フローは、検査対象となる２つのアミノ酸配列の持つ最長共有部分列と、その展開位置を検出するための処理フローである。

【００２５】ＬＣＳ検出部３０は、文字列１で表現されるアミノ酸配列と、文字列２で表現されるアミノ酸配列との間の最長共有文字数を検出する場合には、図３の処理フローに示すように、先ず最初に、ステップ１で、文字列１から１文字ずつ読み込んで、文字列１中での各文字の出現位置を示す表を作成する。

【００２６】この出現表は、例えば、Ａ～Ｚまでのアルファベットに対応した配列に、各文字の出現位置をポインタで連結することによって実現するものであって、例えば、文字列１のアミノ酸配列が“ＡＢＣＢＤＡＢ”で表現される場合には、図６に示すように、“Ａ”が６番目と１番目に出現し、“Ｂ”が７番目と４番目と２番目に出現し、“Ｃ”が３番目に出現し、“Ｄ”が５番目に出現するというように作成する。そして、このステップ１では、更に、以下の処理で用いる文字列１と同じサイズを持つ配列Ｓ[i] の初期化処理を実行して、各エントリにゼロ値を設定する。

【００２７】次に、ステップ２で、文字列２から１文字を読み込み、ステップ１で作成した出現表を参照して、その文字の文字列１での出現位置ｒを特定する。続いて、ステップ３で、用意されている配列Ｓ[i] のｒ番目のＳ[r] のエントリデータと、(r－1)番目のＳ[r－1]のエントリデータとが等しいか否かを判断する。

【００２８】このステップ３で、Ｓ[r] とＳ[r－1]とが等しいと判断するときには、ステップ４に進んで、ｒ番目以上で、かつＳ[r] のエントリデータと等しい値のエントリデータを持つ配列Ｓ[i] に“１”を加算し、続くステップ５で、文字列２の最後の文字までの処理を終了したのか否かを判断して、終了していないことを判断するときには、ステップ２に戻っていく。一方、ステップ３で、Ｓ[r] とＳ[r－1]とが等しくないと判断するときには、ステップ４の加算処理を実行することなく、直ちにステップ５に進んでいく。

【００２９】ここで、ステップ２で読み込んだ文字列２の文字が文字列１で複数回出現する場合には、出現位置ｒの大きい順にステップ３及びステップ４の処理を実行していくことになる。

【００３０】そして、ステップ５で、文字列２の最後の文字までの処理を終了したことを判断すると、ステップ６に進んで、配列Ｓ[i] の最終要素のＳ[m] のエントリデータKmax を最長共有文字数として出力していく。

【００３１】この処理フローの実行により、例えば、文字列１のアミノ酸配列が“ＡＢＣＢＤＡＢ”で表現され、文字列２のアミノ酸配列が“ＢＤＣＡＢＡ”で表現される場合には、文字列２の第１番目の文字Ｂの読込処理に従って、図６の出現表から“ｒ＝７，４，２”が特定されて、図７（ａ）に示すように配列Ｓ[i] のエント

リデータが更新され、文字列２の第２番目の文字Ｄの読込処理に従って、図６の出現表から“ｒ＝５”が特定されて、図７（ｂ）に示すように配列Ｓ[i] のエントリデータが更新され、文字列２の第３番目の文字Ｃの読込処理に従って、“ｒ＝３”が特定されて、図８（ａ）に示すように配列Ｓ[i] のエントリデータが更新され、文字列２の第４番目の文字Ａの読込処理に従って、図６の出現表から“ｒ＝６，１”が特定されて、図８（ｂ）に示すように配列Ｓ[i] のエントリデータが更新される。

【００３２】そして、文字列２の第５番目の文字Ｂの読込処理に従って、図６の出現表から“ｒ＝７，４，２”が特定されて、図９（ａ）に示すように配列Ｓ[i] のエントリデータが更新され、文字列２の第６番目の文字Ａの読込処理に従って、図６の出現表から“ｒ＝６，１”が特定されて、図９（ｂ）に示すように配列Ｓ[i] のエントリデータが更新されていって、最終的に、“４”という最長共有文字数が特定されることになる。なお、図７ないし図９に示す配列Ｓ[i] では、システムの便宜上、文字列１より１文字多いサイズを持つ配列Ｓ[i] に従うもので示してある。

【００３３】次に、図４及び図５に従って、検査対象となる２つのアミノ酸配列の持つ最長共有部分列と、その展開位置を検出するための処理について説明する。ＬＣＳ検出部３０は、文字列１で表現されるアミノ酸配列と、文字列２で表現されるアミノ酸配列との間の最長共有部分列と、その展開位置とを検出する場合には、図４の処理フローに示すように、先ず最初に、ステップ１０で、文字列１から１文字ずつ読み込んで、文字列１中での各文字の出現位置を示す表を作成する。すなわち、図６で説明した出現表を作成するのである。そして、このステップ１では、更に、以下の処理で用いる文字列１と同じサイズを持つ配列Ｓ[i] の初期化処理を実行して、各エントリにゼロ値を設定するとともに、以下の処理で用いる最長共有文字数と同じサイズを持つ配列data[k] の初期化処理を実行して、各エントリが何もポイントしないように設定する。

【００３４】次に、ステップ１１で、文字列２から１文字（ｊ番目の文字）を読み込み、ステップ１０で作成した出現表を参照して、その文字の文字列１での出現位置ｒを特定する。続いて、ステップ１２で、用意されている配列Ｓ[i] のｒ番目のＳ[r] のエントリデータと、(r－1)番目のＳ[r－1]のエントリデータとが等しいか否かを判断する。このステップ１２で、Ｓ[r] とＳ[r－1]とが等しいと判断するときには、ステップ１３に進んで、ｒ番目以上で、かつＳ[r] のエントリデータと等しい値のエントリデータを持つ配列Ｓ[i] に“１”を加算し、一方、等しくないと判断するときには、図５の処理フローのステップ１７の処理に進んで、この加算処理を実行しないよう処理する。ここで、ステップ１１で読み込んだ文字列２の文字が文字列１で複数回出現する場合には、出現位置ｒの大きい順にステップ１２及びステップ１３の処理を実行していくことになる。

【００３５】このようにして設定されるＳ[r] のエントリデータの値ｋが、文字列１のｒ番目の文字までの文字列と、文字列２のｊ番目の文字までの文字列との間の最長共有文字数となる。このように、ＬＣＳ検出部３０は、最長共有部分列を検出していく場合にも、図３の処理フローで説明した最長共有文字数を検出していく処理を実行していくものである。

【００３６】ステップ１３の処理を実行すると、続いて、ステップ１４で、得られたＳ[r]のエントリデータである最長共有文字数ｋに従って、文字列１での展開位置ｒと、文字列２での展開位置ｊとの対データ（ｒ，ｊ）を配列data[k] に格納する。ここで、配列Ｓ[i] が前回の処理サイクルのものから変化していないときには、この格納処理を実行しないように処理する。最長共有部分列は、以下の処理に従って、このデータ構造を連結していくことで求められることになる。

【００３７】続いて、図５の処理フローに移って、ステップ１５で、data[k-1] に格納された文字位置ｒ’，ｊ’を参照して、

ｒ’＜ｒ，　ｊ’＜ｊ

が成立するか否かを判断し、成立すると判断するときには、文字位置の逆転が起こらないことに対応して、ステップ１６に進んで、文字位置ｒ’，ｊ’を次候補となるものとしてポインタを張って登録する。そして、続くステップ１７で、文字列２の最後の文字までの処理を終了したのか否かを判断して、終了していないことを判断するときには、図４の処理フローのステップ１１に戻っていく。一方、ステップ１５で、上述の関係式が成立しないと判断するときには、ステップ１６の処理を実行することなく、直ちにステップ１７の処理に入っていく。

【００３８】そして、ステップ１７で、文字列２の最後の文字までの処理を終了したことを判断すると処理を終了する。この図４及び図５の処理フローの実行により、上述のように、文字列１のアミノ酸配列が“ＡＢＣＢＤＡＢ”で表現され、文字列２のアミノ酸配列が“ＢＤＣＡＢＡ”で表現される場合には、文字列２の第１番目（ｊ＝１）の文字Ｂの読込処理に従って、図６の出現表から“ｒ＝７，４，２”が特定されて、図７（ａ）に示したように、“ｒ＝７”に従って“Ｓ[7] ＝１”が特定されることでdata[1] に（７，１）が格納され、“ｒ＝４”に従って“Ｓ[4] ＝１”が特定されることでdata[1] に（４，１）が格納され、“ｒ＝２”に従って“Ｓ[2] ＝１”が特定されることでdata[1] に（２，１）が格納される。

【００３９】そして、文字列２の第２番目（ｊ＝２）の文字Ｄの読込処理に従って、図６の出現表から“ｒ＝５”が特定されて、図７（ｂ）に示したように、“Ｓ[5] ＝２”が特定されることでdata[2] に（５，２）が

格納される。そして、文字列２の第３番目（ｊ＝３）の文字Ｃの読込処理に従って、図６の出現表から“ｒ＝３”が特定されて、図８（ａ）に示したように、“Ｓ[3]＝２”が特定されることでdata[2]に（３，３）が格納される。そして、文字列２の第４番目（ｊ＝４）の文字Ａの読込処理に従って、図６の出現表から“ｒ＝６，１”が特定されて、図８（ｂ）に示したように、“ｒ＝６”に従って“Ｓ[6]＝３”が特定されることでdata[3]に（６，４）が格納され、“ｒ＝１”に従って“Ｓ[1]＝１”が特定されることでdata[1]に（１，４）が格納される。

【００４０】そして、文字列２の第５番目（ｊ＝５）の文字Ｂの読込処理に従って、図６の出現表から“ｒ＝７，４，２”が特定されて、図９（ａ）に示したように、“ｒ＝７”に従って“Ｓ[7]＝４”が特定されることでdata[4]に（７，５）が格納され、“ｒ＝４”に従って“Ｓ[4]＝３”が特定されることでdata[3]に（４，５）が格納され、“ｒ＝２”に従って“Ｓ[2]＝２”が特定されることでdata[2]に（２，５）が格納される。そして、文字列２の第６番目（ｊ＝６）の文字Ａの読込処理に従って、図６の出現表から“ｒ＝６，１”が特定されて、図９（ｂ）に示したように、“ｒ＝６”に従って“Ｓ[6]＝４”が特定されることでdata[4]に（６，６）が格納される。なお、図９（ｂ）から分かるように、“ｒ＝６”と“ｒ＝１”とで配列Ｓ[i]に変化がないことから、（１，６）の格納処理は実行されない。

【００４１】そして、これらの文字位置情報（文字列１と文字列２の持つ同一文字の展開位置を表示する）は、data[k-1]に格納されたものと、data[k]に格納されたものとで文字位置の逆転が起こらない場合には、それらの間でポインタが張られていくことで、図１０のように、data[k]に格納されるのである。

【００４２】最長共有部分列は、このdata[k]に格納される文字位置情報のポインタを辿っていくことで特定されることになる。すなわち、図１０の例で説明するならば、“data[4]の（７，５）→data[3]の（６，４）→data[2]の（５，２）→data[1]の（４，１）”という連結に従って、最長共有部分列ＢＤＡＢと、文字列１及び文字列２におけるその展開位置が特定され、“data[4]の（７，５）→data[3]の（６，４）→data[2]の（５，２）→data[1]の（２，１）”という連結に従って、最長共有部分列ＢＤＡＢと、文字列１及び文字列２におけるその展開位置が特定され、“data[4]の（７，５）→data[3]の（６，４）→data[2]の（３，３）→data[1]の（２，１）”という連結に従って、最長共有部分列ＢＣＡＢと、文字列１及び文字列２におけるその展開位置が特定され、“data[4]の（６，６）→data[3]の（４，５）→data[2]の（３，３）→data[1]の（２，１）”という連結に従って、最長共有部分列ＢＣＢＡと、文字列１及び文字列２におけるその展開位置が特定されるのである。

【００４３】図１１及び図１２に、ＬＣＳ検出部３０が、この連結を辿っていくことで最長共有部分列を特定していくときに実行する処理フローを図示する。次に、図２に示した遺伝子情報検査装置１の各処理部が、このＬＣＳ検出部３０の検出する最長共有文字数と、最長共有部分列及びその展開位置とを受けて実行することになる処理について説明する。

【００４４】相同性判定部３１は、ＬＣＳ検出部３０が入力装置４０から入力されてくるアミノ酸配列の文字列（以下、入力アミノ酸配列と称する）と、アミノ酸配列データベース５０やモチーフデータベース６０から与えられるアミノ酸配列の文字列との間の最長共有文字数を検出すると、その最長共有文字数と入力アミノ酸配列の文字列長との割合値を検出して、その割合値が規定の基準値よりも大きい場合には、入力アミノ酸配列が、アミノ酸配列データベース５０やモチーフデータベース６０から与えられるアミノ酸配列と相同であると判定し、基準値よりも小さい場合には、相同でないと判定する。

【００４５】相同性探索部３２は、相同性判定部３１の判定結果を利用して、入力アミノ酸配列と相同なアミノ酸配列をアミノ酸配列データベース５０から検索する。そして、相同の関係にある場合には、相同性判定部３１により算出された割合値と、ＬＣＳ検出部３０により検出された最長共有部分列とを表示部３５を介してディスプレイ装置４２に表示する。

【００４６】図１３に、この表示例の一例を図示する。この表示例は、ヒトチトクロームｃとバクテリアチトクロームｃという２つのアミノ酸配列の処理結果を表示するものであって、最長共有部分列については、両者のアミノ酸配列にどのような文字間隔でもって配置されているかを示す表示形態に従って表示する構成を採っている。すなわち、“ＧＤ｛ｘ３，３｝Ｇ｛ｘ０，１｝Ｋ｛ｘ０，２｝・・”と表示する形態を採って、ヒトチトクロームｃでは、“ＧＤ”の後３文字については一致しない文字が続いて、その後に“Ｇ”が続いて、その後直ぐに“Ｋ”が続き、一方、バクテリアチトクロームｃでは、“ＧＤ”の後３文字については一致しない文字が続いて、その後に“Ｇ”が続いて、その後１文字については一致しない文字が続いて、その後に“Ｋ”が続くというように表示するものである。

【００４７】モチーフ探索部３３は、相同性判定部３１の判定結果を利用して、先ず最初に、入力アミノ酸配列と相同なモチーフ配列をモチーフデータベース６０から検索し、続いて、ＬＣＳ検出部３０の検出する最長共有部分列と、その配列位置間の持つ文字列長とに従って、この相同の関係にあるモチーフ配列が本来のモチーフ配列であるか否かを判定する。例えば、“Ｌ”の後に規定されない文字が６文字続いてその後に“Ｌ”が続き、こ

の“Ｌ”の総個数が５個となるロイシンジッパーというモチーフ配列が、規定されない６文字まで含めて入力アミノ酸配列に含まれているか否かをＬＣＳ検出部３０の検出結果に従ってチェックしていくのである。そして、モチーフ探索部３３は、入力アミノ酸配列がモチーフ配列を持つ場合には、入力アミノ酸配列とモチーフ配列とを表示部３５を介してディスプレイ装置４２に表示する。図１４に、ロイシンジッパーを持つラット卵細胞カリウムチャンネルの表示例を図示する。

【００４８】アラインメント部３４は、ＬＣＳ検出部３０の検出する最長共有部分列と、その展開位置とを受けて、入力アミノ酸配列と、アミノ酸配列データベース５０やモチーフデータベース６０から与えられるアミノ酸配列の持つ最長共有部分列が対応付けられるべく、この２つのアミノ酸配列をアラインメントして表示部３５を介してディスプレイ装置４２に表示する。図１５に、この表示例の一例を図示する。この表示例は、ヒトチトクロームｃとバクテリアチトクロームｃという２つのアミノ酸配列の処理結果を表示するものであって、配列位置間の持つ文字列長分に相当する空白を挿入していくことでアラインメント処理を実行していくことになる。

【００４９】

【発明の効果】以上説明したように、本発明によれば、文字列で表現される検査対象アミノ酸配列と、文字列で表現される比較対象アミノ酸配列との最長共有文字数や最長共有部分列を検出する構成を採って、この最長共有文字数や最長共有部分列に従って、検査対象アミノ酸配列と比較対象アミノ酸配列との間の類似性を評価する構成を採るものであることから、ダイナミックプログラミング手法によるアミノ酸配列の比較方法に比べて、小さなメモリ容量で、かつ高速に２つのアミノ酸配列の類似性を評価することができるのである。

【図面の簡単な説明】

＊【図１】本発明の原理構成図である。

【図２】本発明の一実施例である。

【図３】ＬＣＳ検出部の実行する処理フローの一実施例である。

【図４】ＬＣＳ検出部の実行する処理フローの一実施例である。

【図５】ＬＣＳ検出部の実行する処理フローの一実施例である。

【図６】ＬＣＳ検出部の作成する出現表の説明図である。

【図７】配列Ｓ[i]の更新処理の説明図である。

【図８】配列Ｓ[i]の更新処理の説明図である。

【図９】配列Ｓ[i]の更新処理の説明図である。

【図１０】ＬＣＳ検出部の作成するデータ構造の説明図である。

【図１１】ＬＣＳ検出部の実行する処理フローの一実施例である。

【図１２】ＬＣＳ検出部の実行する処理フローの一実施例である。

【図１３】処理結果の表示形態の一実施例である。

【図１４】処理結果の表示形態の一実施例である。

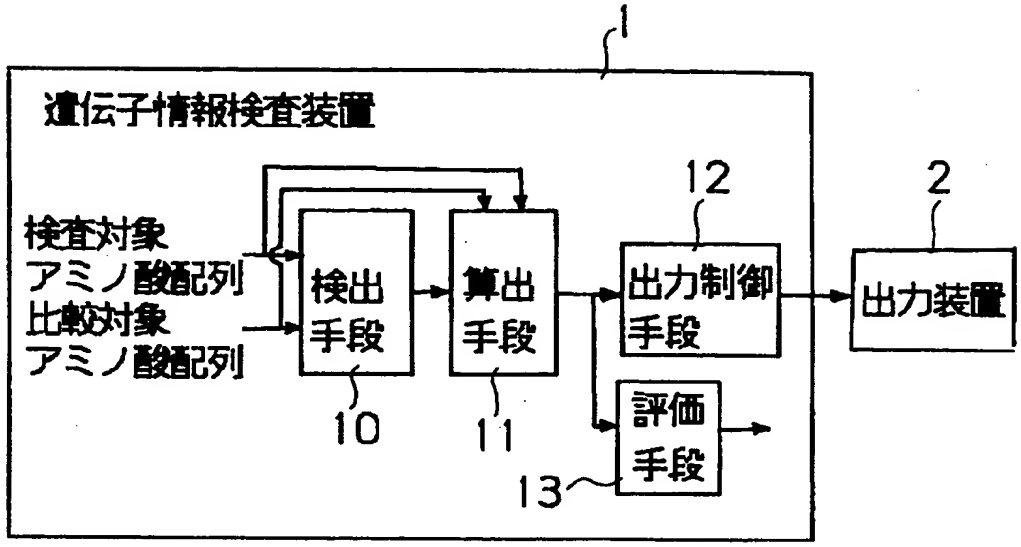
【図１５】処理結果の表示形態の一実施例である。

【符号の説明】

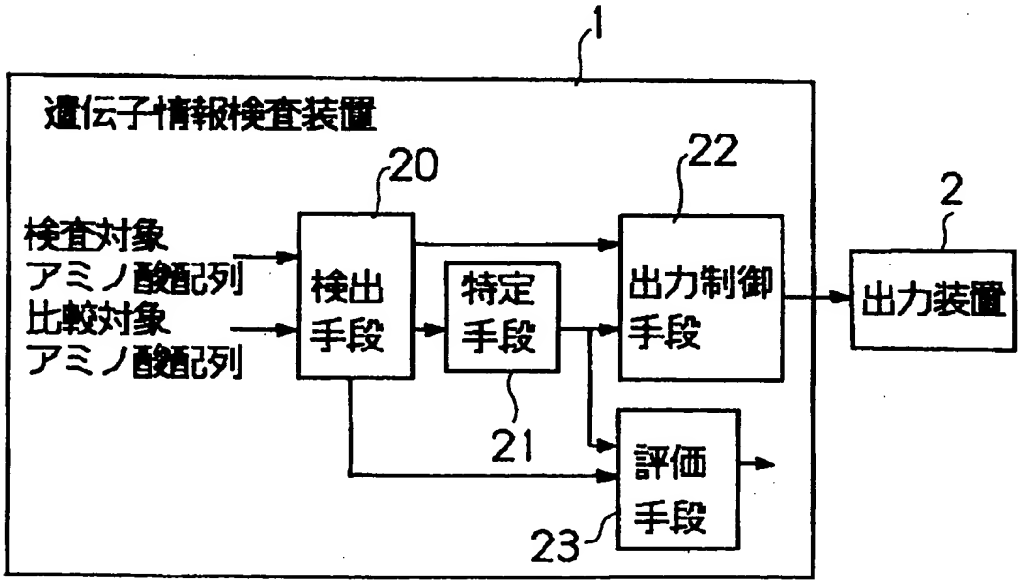
１　遺伝子情報検査装置
２　出力装置
１０　検出手段
１１　算出手段
１２　出力制御手段
１３　評価手段
２０　検出手段
２１　特定手段
２２　出力制御手段
２３　評価手段

【図１】

本発明の原理構成図



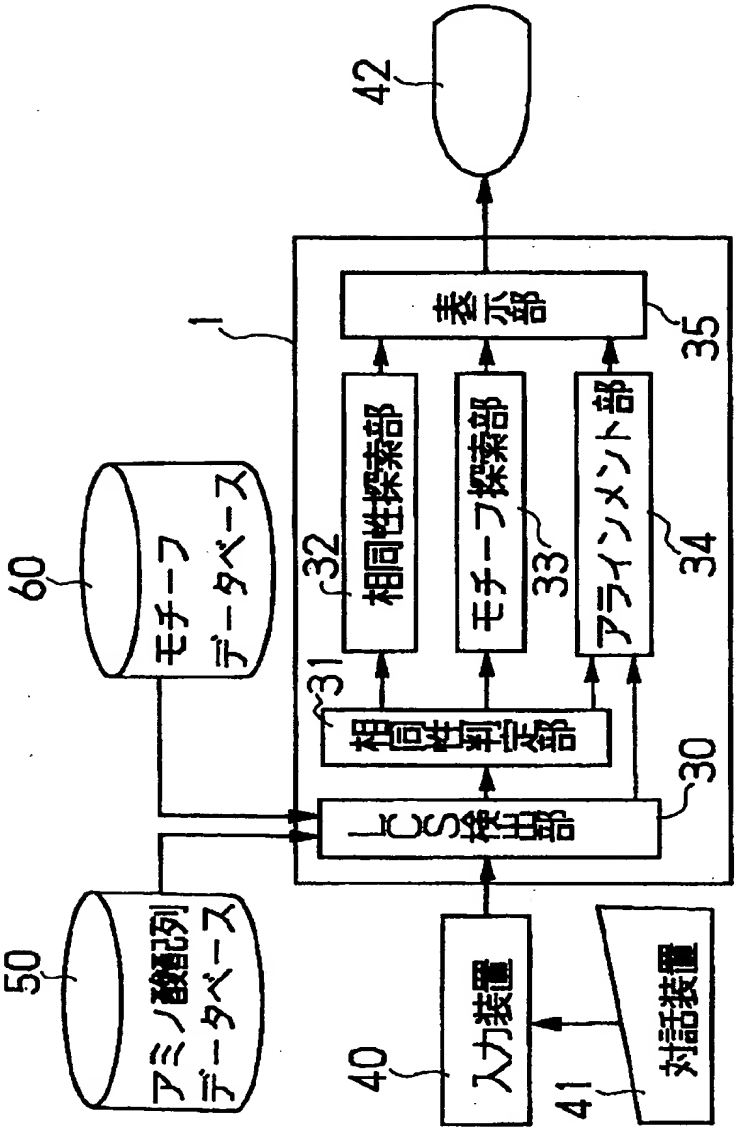
(a)



(b)

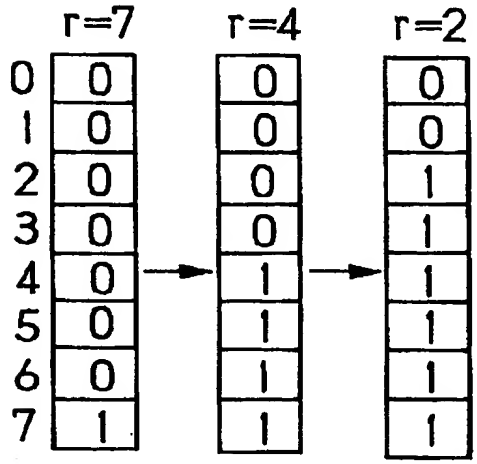
【図２】

本発明の一実施例

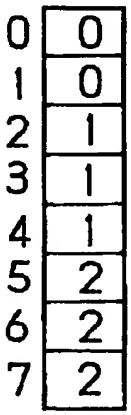


【図７】

配列S[i]の更新処理の説明図



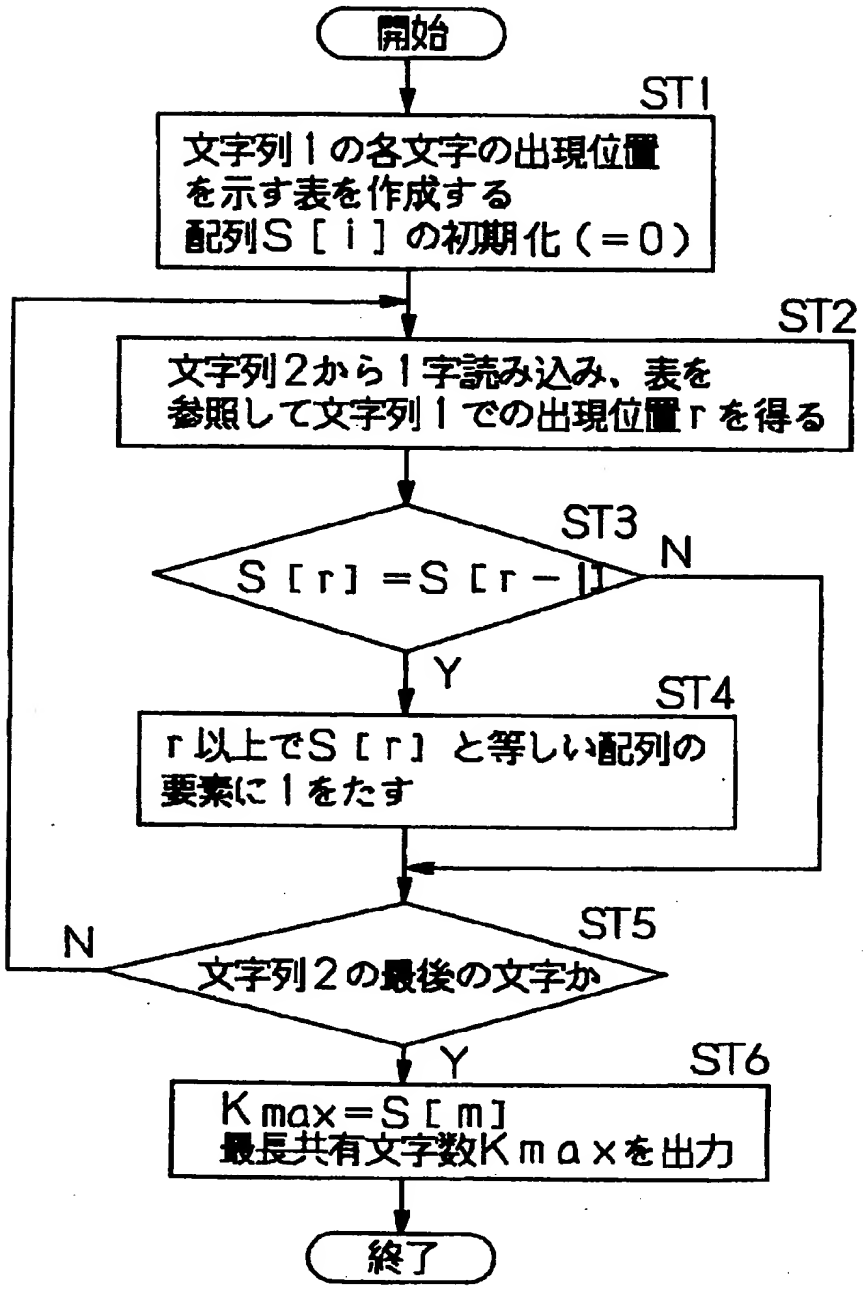
(a)



(b)

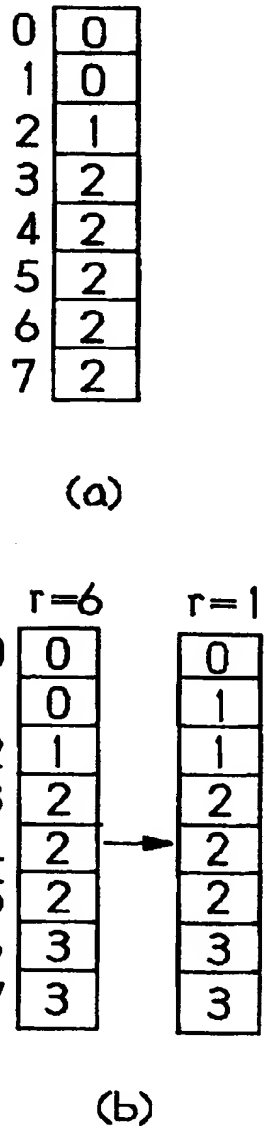
【図３】

ＬＣＳ検出部の実行する処理フローの一実施例



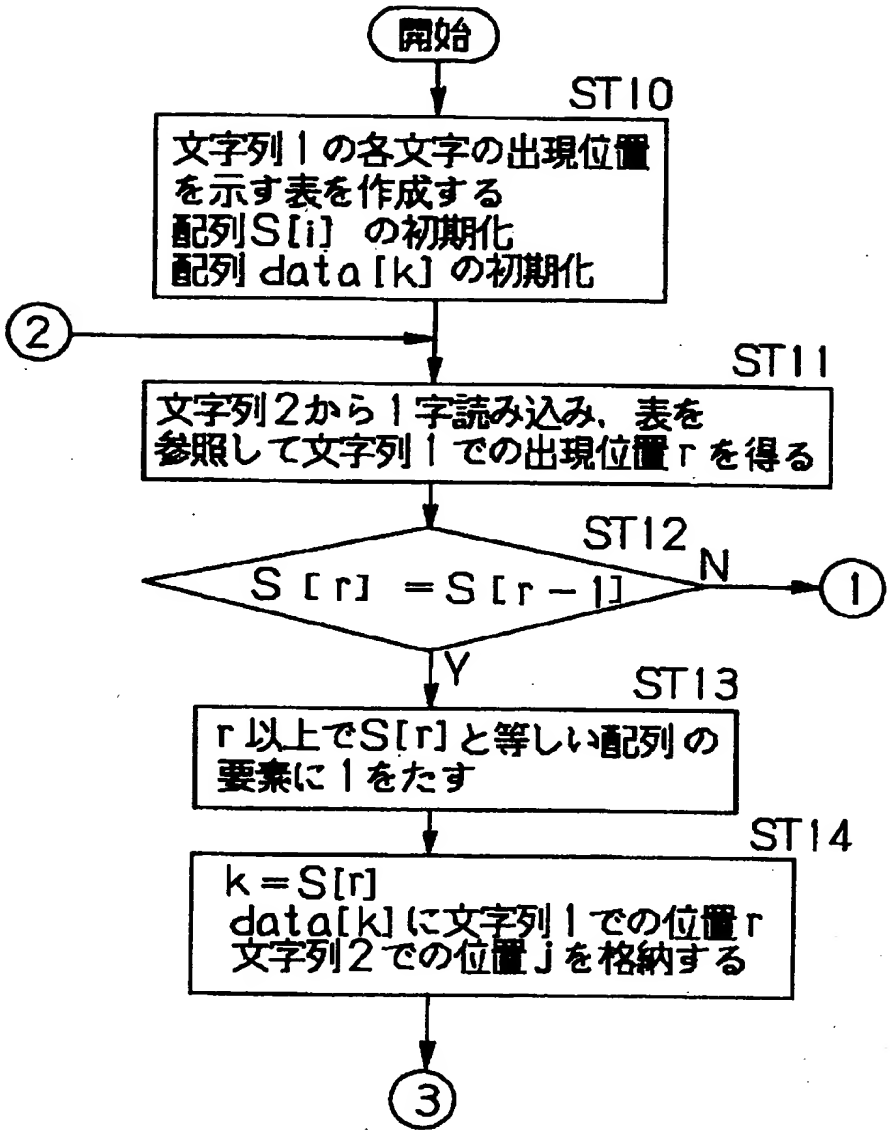
【図８】

配列Ｓ[i]の更新処理の説明図



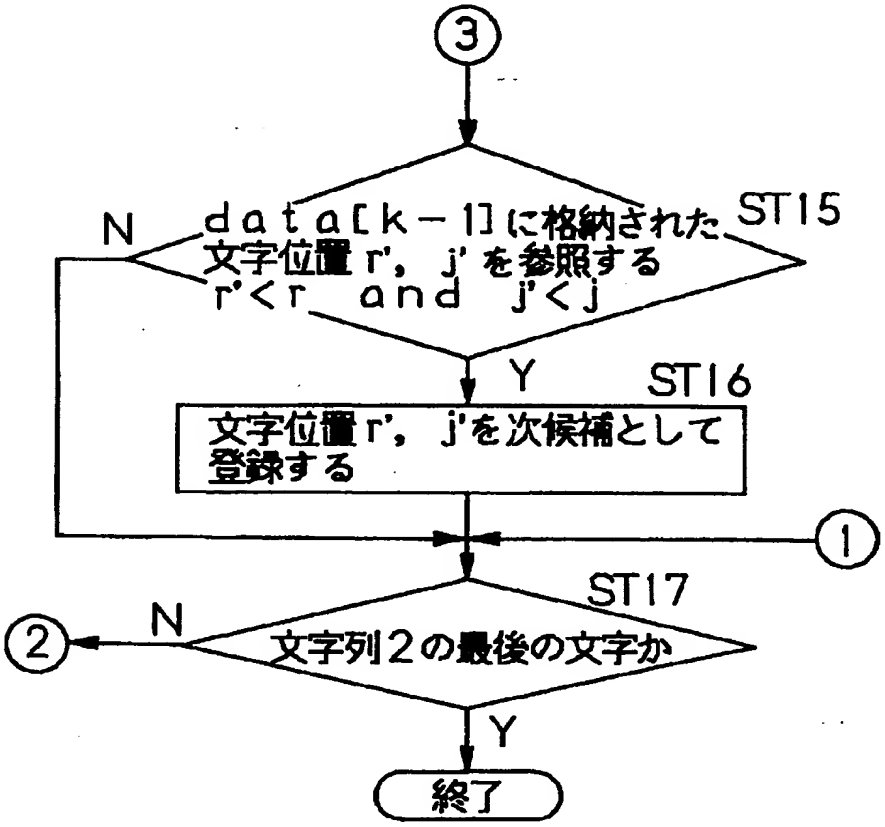
【図４】

ＬＣＳ検出部の実行する処理フロー の一実施例



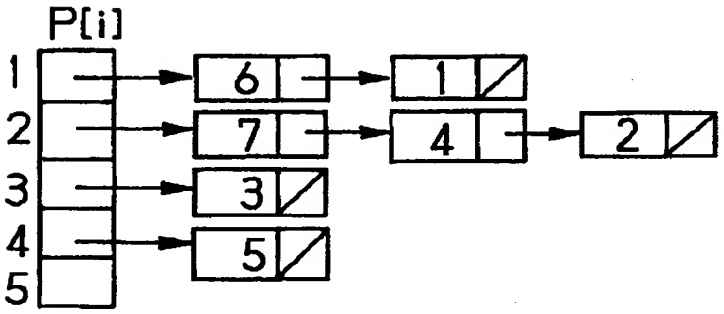
【図５】

ＬＣＳ検出部の実行する処理フローの一実施例



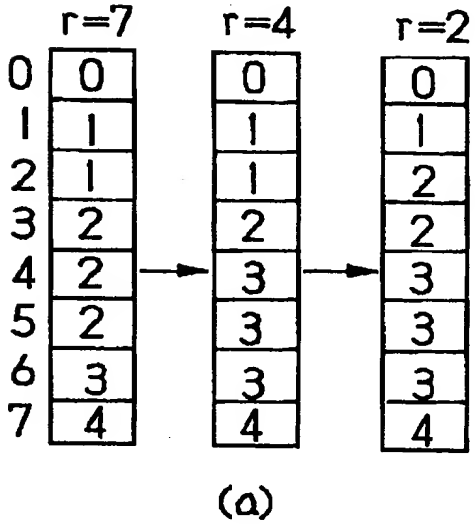
【図６】

ＬＣＳ検出部の作成する出現表の説明図

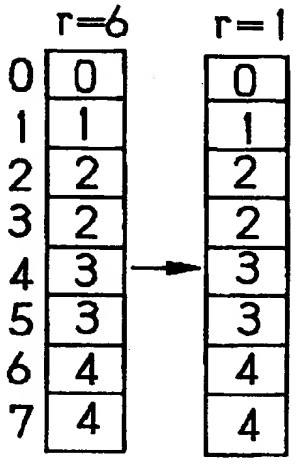


【図９】

配列S[i]の更新処理の説明図



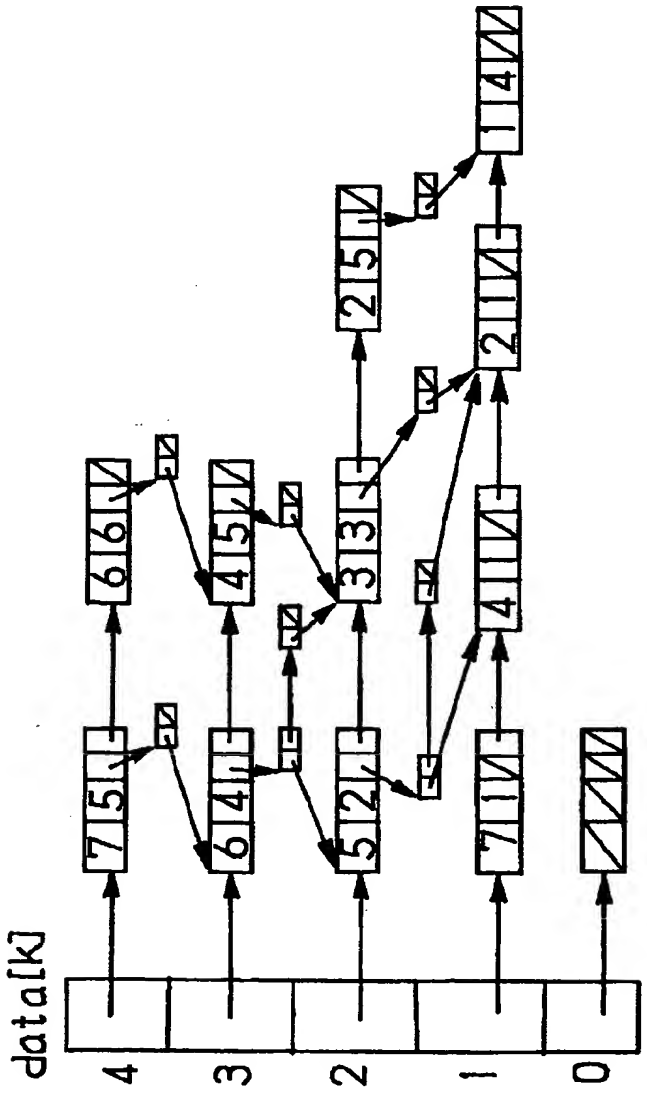
(a)



(b)

【図10】

LCS検出部の作成するデータ構造の説明図



【図14】

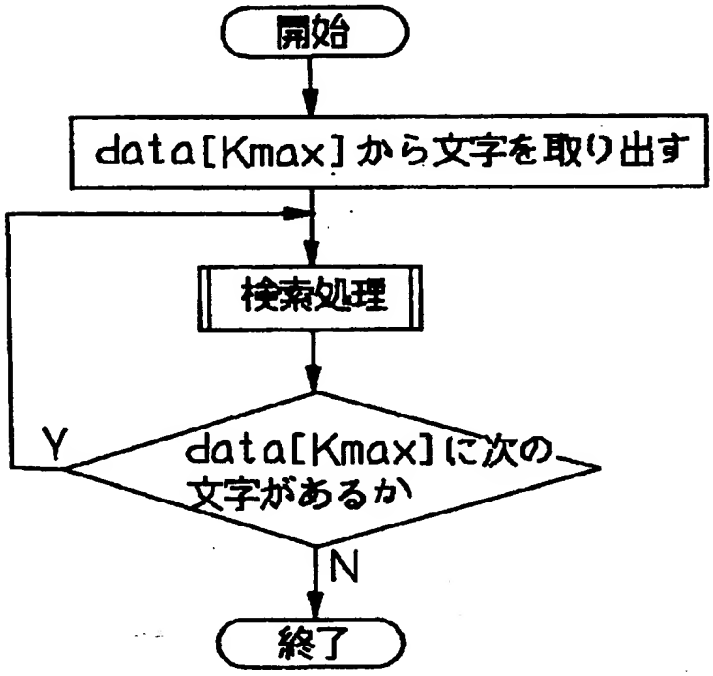
処理結果の表示形態の一実施例

Rat　:MSLAILRVIRLVRVFRIFKLSRHSKGLQILGRTLKASMRELGLLIFFIFIGVV...

leucinzip: L(6)L(6)L(6)L(6)L

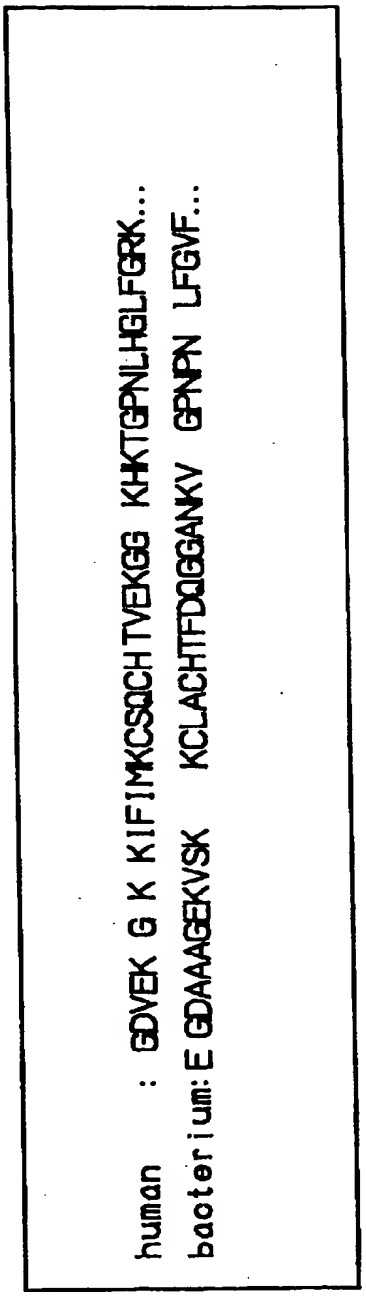
【図１１】

ＬＣＳ検出部の実行する処理フローの 一実施例



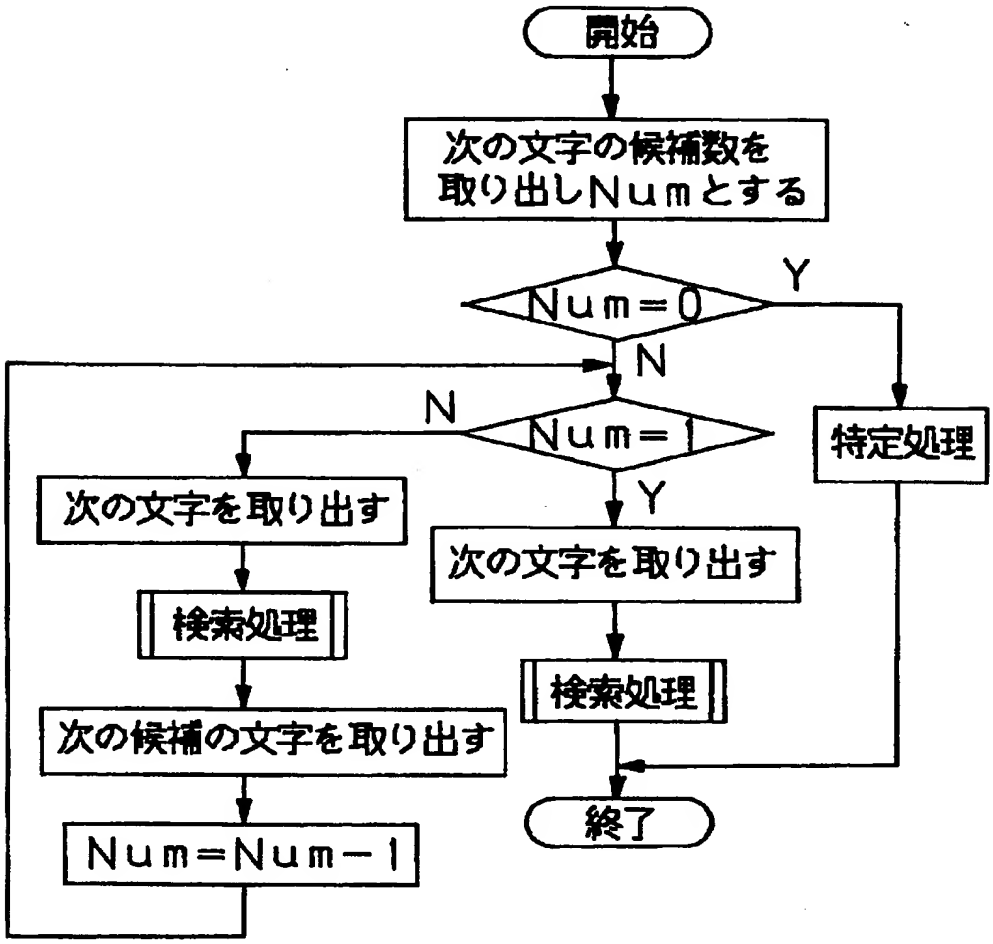
【図１５】

処理結果の表示形態の一実施例



【図１２】

ＬＣＳ検出部の実行する処理フローの一実施例



【図１３】

処理結果の表示形態の一実施例